Nonparametric Probability Density Estimation for Data Analysis in Several Dimensions[1]

David W. Scott

Rice University
Houston, TX 77251

1. Introduction

Our purpose in this paper is to illustrate how nonparametric probability density estimates, in particular the corresponding contour curves, are a useful adjunct to scatter diagrams when performing a preliminary examination of a set of random data in several dimensions. For a preliminary approach we generally want to perform fairly simple tasks with free-form techniques to uncover structures and features of interest in the data. Such procedures are often graphical and unlike summary statistics seldom lead to much compression of the data. Tukey (1977) presents a wealth of such procedures. One which well illustrates the power and flexibility of these preliminary procedures is the running median smoothing algorithm for time series data (with resmoothing of the rough and the like). Other graphical techniques for multivariate data are presented in Tukey and Tukey (1981).

For preliminary viewing of one-dimensional data, both scatter diagrams and frequency curves such as histograms are widely and successfully employed to examine clustering, tail behavior, and skewness of

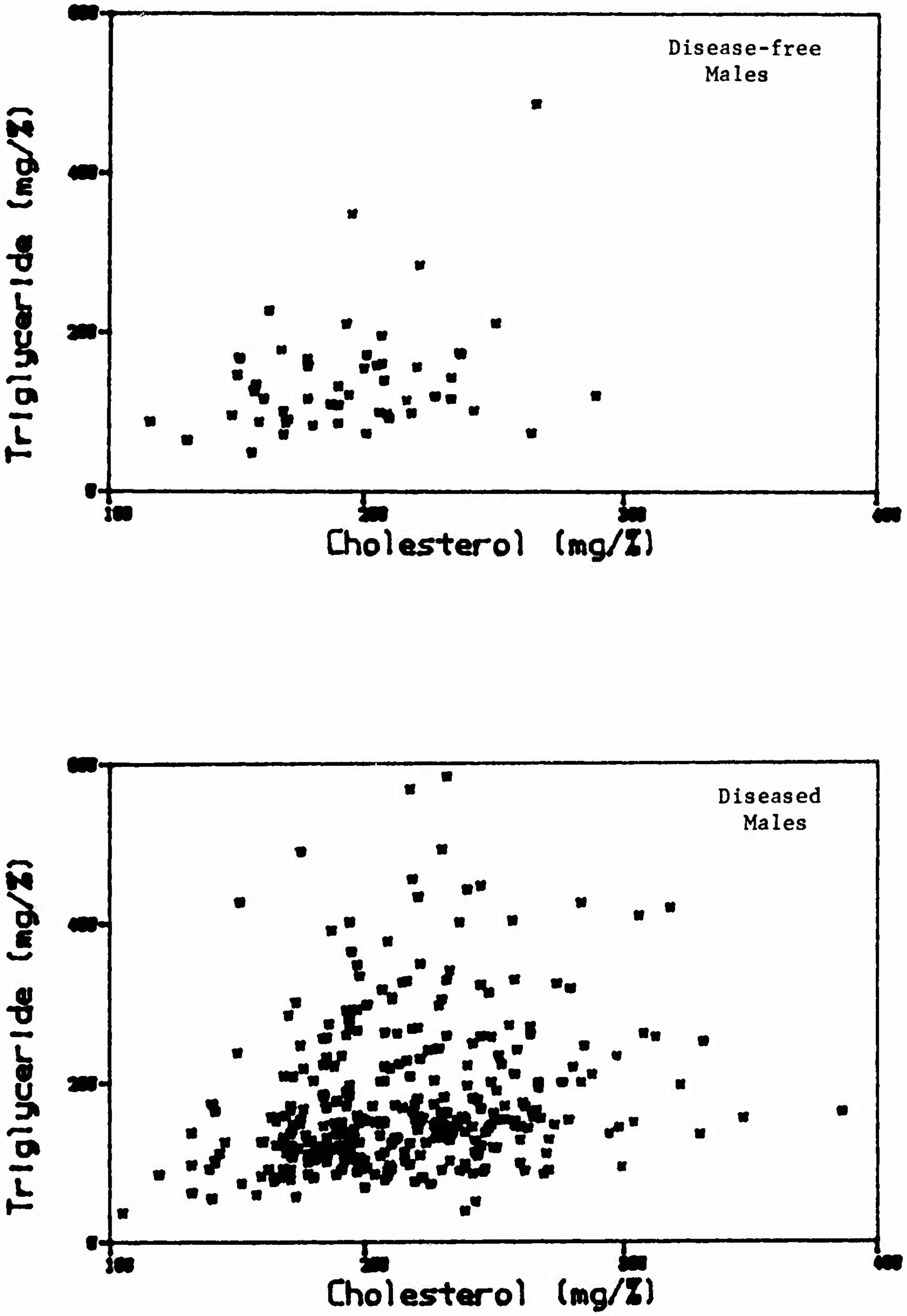
[1] This research was supported in part by the Army Research Office under DAAG-29-82-K-0014 and by NASA/Lockheed under PO-0200100079.

data. For bivariate data, scatter diagrams are in practice widely preferred to bivariate frequency curves. Scatter diagrams of three dimensional data may be realized by viewing a projection of the data on a rotating plane represented by the screen on a computer graphics terminal. For higher dimensions carefully selected projections may also be viewed, and sophisticated techniques have been developed, and are evolving, for choosing good projections (Friedman and Tukey, 1974). Apparently the success of frequency curves in one dimension has not readily extended to higher dimensions. It is an open question as to the number of dimensions that may be successfully visualized with a nonparametric density estimator under various conditions (sample size, for example). It is our purpose to illustrate the power of preliminary frequency curves as an adjunct to viewing scatter diagrams.

2. Bivariate Data

We shall examine a data set which contains information on the status of the coronary arteries of 371 men suspected of having heart disease, having experienced episodes of severe chest pain. These data have been more fully described and analyzed; see Gotto, *et al.* (1977) and Scott, *et al.* (1978). After visual examination of the coronary arteries by angiography, 51 men were determined to be free of significant coronary artery disease. It was of interest to compare the levels of blood fats, plasma cholesterol and plasma triglyceride concentrations, between the group of 51 disease-free males and the group of 320 diseased males. The scatter diagrams of these two data sets are displayed in Figure 1. Patients with elevated levels of cholesterol and

Figure 1. Scatter Diagrams



triglyceride are evident among the diseased males. This observation is difficult to evaluate in light of the large difference in sample sizes. However, it is unlikely that a larger sample of 320 disease-free males would result in a scatter diagram similar to that of the 320 diseased males.

To obtain a nonparametric density contour plot we computed a bivariate product kernel estimate (Epanechnikov, 1969) given by

$$f(x,y) = \frac{1}{nh_x h_y} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h_x}\right) K\left(\frac{y_i - y}{h_y}\right) \qquad (1)$$

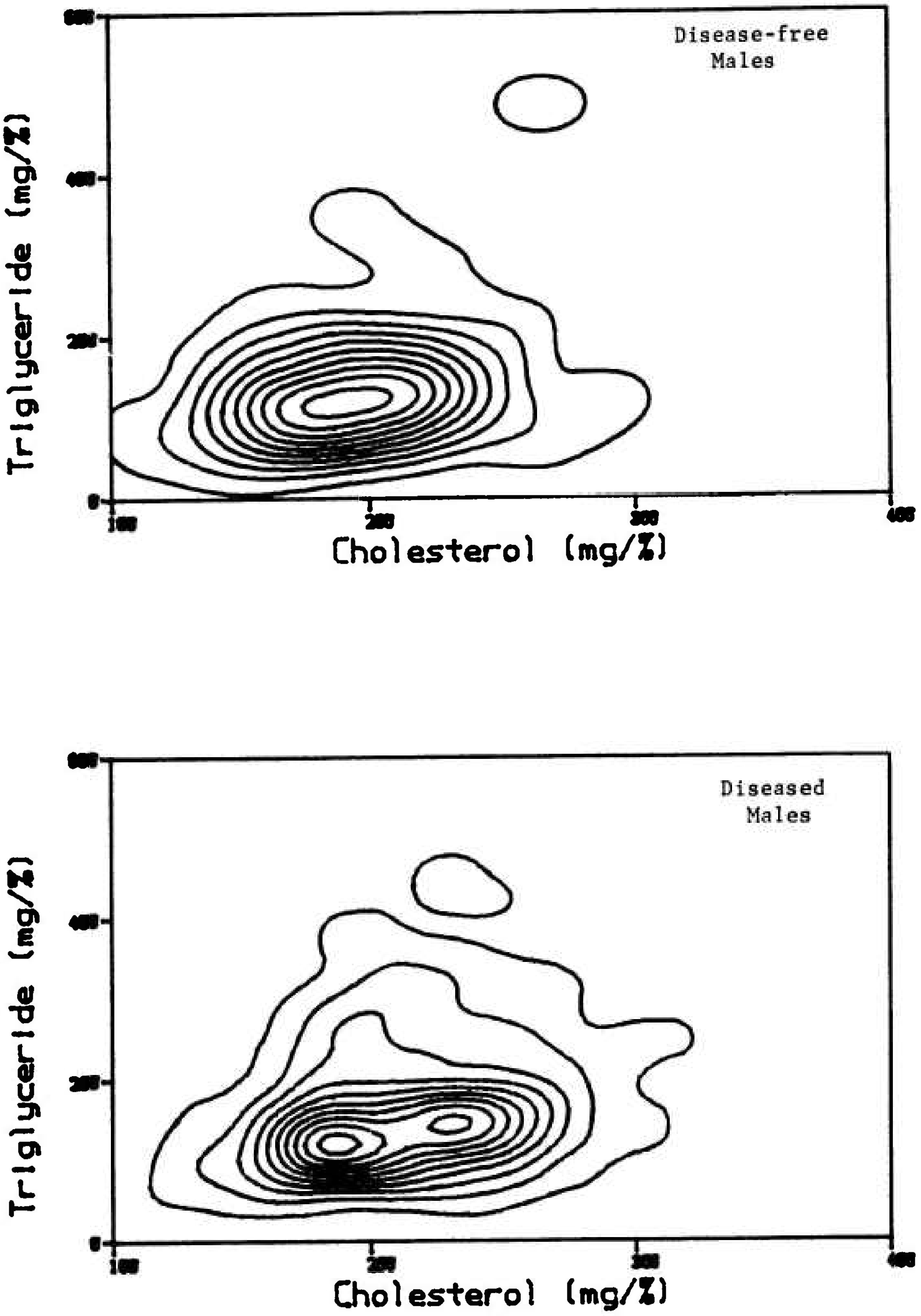
using a quartic (biweight) kernel

$$K(z) = \frac{15}{16}(1-z^2)^2 \, I_{[-1,1]}(z) \qquad (2)$$

and preliminary values of the smoothing parameters given by $h_x = 2 s_x n^{-1/6}$ where s_x represents a trimmed and pooled estimate of the standard deviation for the two groups with a similar expression for h_y. Density values were computed over a grid of 150 by 90 points. When applied to the data for the diseased males, the contour plot reveals a striking bimodal feature, as shown in Figure 2. The contours of equal probability are at the ten levels 0.05 to 0.95 in increments of 0.10 as a fraction of the respective maximal modal levels. The density function of the disease-free males could be well approximated by a bivariate Normal form. Its mode coincides with the left of the two modes in the density function of the diseased males.

The contour plots have helped emphasize a feature in the scatter diagram that might have gone unnoticed. The contour plots also aid in compensating for the difference in sample sizes. The discovery of the bimodal feature led to formulation of a complex cholesterol-triglyceride

Figure 2. Bivariate Density Contours



interaction in the model for estimating the risk of coronary artery disease. Clinically, the difference of 50 mg/% between the two modes in Figure 2 for the diseased males is greater than the reduction in cholesterol by dietary intervention (which usually achieves proportional reductions in the range of 10 to 15 percent).

3. Trivariate Data

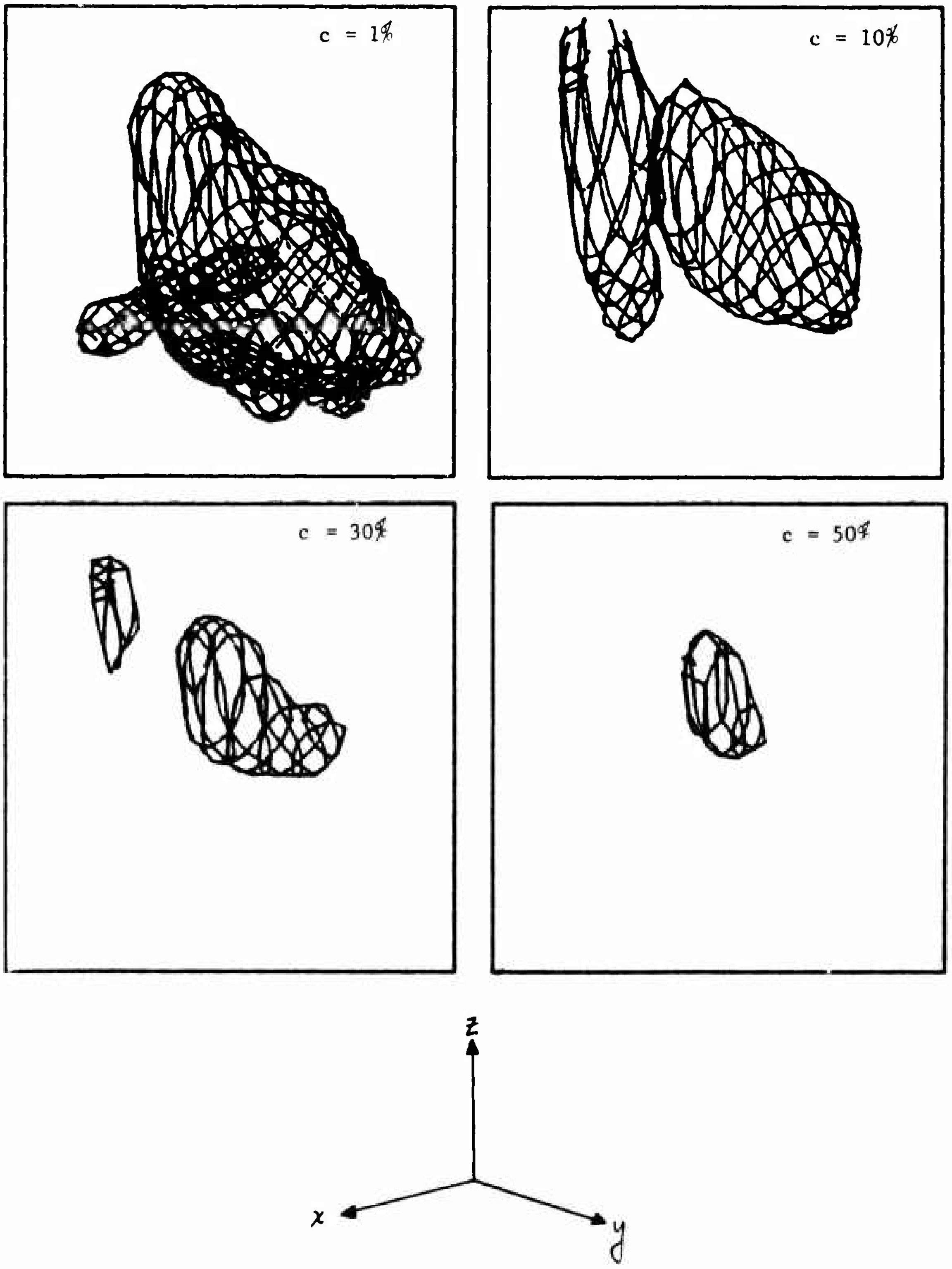
The data presented in this section were obtained by processing four-channel Landsat data measured over North Dakota during the summer growing season of 1977 and were furnished by Dick Heydorn of NASA/Houston and Chuck Sorensen of Lockheed/Houston. The sample contains approximately 21,000 points, each representing a 1.1 acre pixel, covering a 5 by 6 nautical mile section. On each pass over an individual pixel by the Landsat satellite, the four channel readings were combined into a single value that measures the "greenness" of the pixel at that time. The greenness of a pixel was plotted as a function of time from the five passes during the growing season. Finally, Badhwar's (1982) growth model was fitted to this curve. This model has three parameters which are contained in each trivariate data point. The first variable (x) gives the time the "crop" (if any) ripened. The second variable (y) measures the approximate time to ripen. And the third variable (z) measures the level of "greenness" at the time of ripening. Although it is natural to group these data by actual type of ground cover for classification procedures, we have not done so here.

It is not possible to present a satisfactory picture of a three-dimensional scatter diagram of these data for this article. However, on an AED512 terminal with 512 by 512 resolution, a projection of these data onto the screen typically displayed only 4000 points, the rest being "hidden" behind displayed points. Viewed from several different angles, various shapes and features in the data were easily perceived. Color was used to indicate the level of the variable perpendicular to the screen.

We can present density contours of an estimate f(x,y,z). Consider an equiprobable contour at level c; that is, consider those points (x,y,z) satisfying the equation f(x,y,z) = c. The solution of this equation for a smooth density estimate f is a smooth surface (or surfaces) in $\mathbb{R}^3$. This surface may be displayed by intersecting it with a series of planes displaced equal distances along the co-ordinate axes, in the following, along only the x and y axes. In Figure 3, we display the surface for c = 1% of the maximal mode value. Comparing Figure 3 to the corresponding scatter diagram on the same projection plane reveals how surprisingly little of the data space is enclosed in this contour. In the scatter diagram our eyes focused on rays of points that seemed interesting but represented only a small fraction of the data. Also notable in Figure 3 is a cylindrical shape disjoint and behind the larger surface. This feature was also clearly visible in the scatter diagram and represents acres in which sugar beets were grown. Apparently the method by which sugar beets are harvested leads to a singularity in the estimation of the growth model parameters with $y \approx 0$.

Expanding the scale by a factor of 2 while retaining the same center as in the c = 1% picture, we show the contour shapes at levels c = 10%, 30%, and 50% of modal height. Notice how each contour shape

Figure 3. Trivariate Density Contours



"fits" inside the preceding one. Also observe how multimodal features appear in this space. Three modes are shown in this sequence. On a color graphics terminal, we may simultaneously view these and other contours by using different colors to draw each contour.

Again, the density plots have complemented and added to our understanding of these data. It is easier to see inside the data cloud with this representation and also makes rotation of the data cloud less important.

4. Computational Considerations

A new algorithm and density estimator were developed to display the trivariate contour plots and we hope to report on it in another paper (Scott, 1983b). Speed is an important factor in an interactive environment. The kernel method used in the bivariate case becomes excruciatingly slow when presented with 21,000 points in three dimensions. In real time, a few minutes were required on a Vax 11/780 to compute the bivariate kernel contours for 320 points on a 150 by 90 mesh. To generate the pictures in Figure 3, we evaluated the density on a 30 by 30 by 30 mesh for 21,000 points. A straightforward kernel estimator would have required several hours to compute!

The histogram estimator is extremely efficient computationally, but very inefficient statistically -- and relatively more inefficient in higher dimensions than kernel methods. One recent discovery indicates that the frequency polygon may be a good choice of a nonparametric density estimator since it is computationally equivalent to a histogram but statistically similar to a kernel estimate (Scott, 1983a). However, the frequency polygon in several dimensions suffers from sensitivity to choice of cell boundaries. The new algorithm addresses this problem and is asymptotically equivalent to a certain kernel estimate. Other fast preliminary estimates in one and two dimensions may be obtained by numerical approximation of kernel estimates in place of statistical approximation, which we prefer.

5. Where Do We Go?

We do not really know for how many dimensions nonparametric density estimates will be useful and feasible. Scatter diagrams have been used in a highly interactive environment to visualize nine-dimensional data (Tukey, Friedman, and Fisherkeller, 1976). Many possible strategies may be envisioned for using color and motion to examine data in more than three dimensions. We expect much progress in this area. But for larger and larger data sets requiring sophisticated analysis, we believe that density-based methods will be both efficient and effective.

REFERENCES

Badhwar, G.D., J.G. Carnes, and W.W. Austin (1982), "Use of Landsat-Derived Temporal Profiles for Corn-Soybean Feature Extraction and Classification," *Remote Sensing of Environment*, 12, 57-79.

Epanechnikov, V.A. (1969), "Nonparametric Estimates of a Multivariate Probability Density," *Th. Prob. and Appl.*, 14, 153-158.

Friedman, J.H. and J.W. Tukey (1974), "A Projection Pursuit Algorithm for Exploratory Data Analysis," *IEEE Trans. Comp. C-23*, 881-890.

Gotto, A.M., G.A. Gorry, J.R. Thompson, J.S. Cole, R. Trost, D. Yeshurun, M.E. DeBakey (1977), "Relationship Between Plasma Lipid Concentration and Coronary Artery Disease in 496 Patients," *Circulation*, 56:5, 875-883.

Scott, D.W. (1983a), "Optimal Frequency Polygons: Theory and Application," submitted.

Scott, D.W. (1983b), "Average Shifted Histograms," working paper.

Scott, D.W., A.M. Gotto, J.S. Cole, and G.A. Gorry (1978), "Plasma Lipids as Collateral Risk Factors in Coronary Artery Disease - A Study of 371 Males with Chest Pain," *J. Chronic Diseases*, 31, 337-345.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Tukey, J.W., J.H. Friedman, and M.A. Fisherkeller (1976), "PRIM-9, an Interactive Multidimensional Data Display and Analysis System," *Proc. 4th Inter. Congress for Stereology*, Sept. 4-9, 1975, Gaithersburg, Maryland.

Tukey, P.A. and J.W. Tukey (1981), "Graphical Display of Data Sets in 3 or More Dimensions," in *Interpreting Multivariate Data*, V. Barnett, ed., John Wiley & Sons, New York.